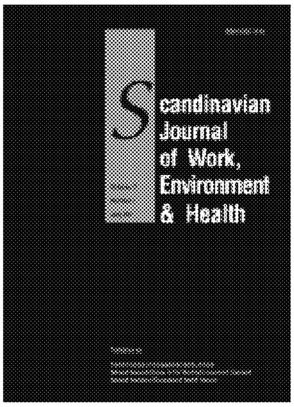

Scandinavian Journal of Work, Environment & Health

Original article

Scand J Work Environ Health 2000;26(1):37-43

Biases in estimating the effect of cumulative exposure in log-linear models when estimated exposure levels are assigned

by Steenland K, Deddens JA, Zhao S

Key terms: cumulative exposure; measurement error; occupation

Scand J Work Environ Health 2000;26(1):37—43

Biases in estimating the effect of cumulative exposure in log-linear models when estimated exposure levels are assigned

by Kyle Steenland, PhD,[1] James A Deddens, PhD,[1,2] Shuhong Zhao, MA[2]

Steenland K, Deddens JA, Zhao S. Biases in estimating the effect of cumulative exposure in log-linear models when estimated exposure levels are assigned. *Scand J Work Environ Health* 2000;26(1):37—43.

Objectives Exposure-response trends in occupational studies of chronic disease are often modeled via log-linear models with cumulative exposure as the metric of interest. Exposure levels for most subjects are often unknown, but can be estimated by assigning known job-specific mean exposure levels from a sample of workers to all workers. Such assignment results in (nondifferential) measurement error of the Berkson type, which does not bias the estimate of exposure effect in linear models but can result in substantial bias in log-linear models with dichotomous outcomes. This bias was explored in estimated exposure-response trends using cumulative exposure.
Methods Simulations were conducted under the assumptions that (i) exposure level is assigned to all workers based on the job-specific means from a sample of workers, (ii) exposure level and duration are log-normal, (iii) the true exposure-response model is log-linear for cumulative exposure, (iv) the disease is rare, and (v) the variance of job-specific exposure level increases with its job-specific mean.
Results Assignment of job-specific mean exposure levels from a sample of workers causes an upward bias in the estimated exposure-response trend when there is little variance in the duration of exposure but causes a downward bias when duration has a large variance. This bias can be substantial (eg, 30—50%).
Conclusions Berkson errors in exposure result in little bias in estimating exposure-response trends when the standard deviation of duration is approximately equal to its mean, which is common in many occupational studies. No bias occurs when the variance of exposure level is constant across jobs, but such conditions are probably uncommon.

Key terms cumulative exposure, measurement error, occupation.

Cumulative exposure is often used as the exposure metric in modeling exposure-response trends for chronic disease. In many studies the exposure level for most or all subjects is not measured but is instead assigned based on measurements of a sample of subjects. In many occupational studies, for example, most workers may have been exposed in the past when no measurements were taken. Current measurements can be made on a sample of current workers in a variety of jobs, and the job-specific averages from the current measurements can then be assigned to all the workers in these jobs, including past workers, as long as exposures can be assumed to have remained constant over time. If changes in exposure level are likely over time, then the changes can be estimated and the assigned exposure levels can be adjusted accordingly.

Assigning exposure level using the means for specific jobs necessarily results in (nondifferential) error in the assigned exposure level compared with the true exposure. The resulting errors conform to a Berkson error model, in which the true exposure level in a given job category can be assumed to vary randomly about the assigned level or "observed" level (1, 2),

$$\text{exposure}_{\text{true}} = \text{exposure}_{\text{observed}} + \varepsilon.$$

In this case the error is independent of the observed exposure and is, consequently, positively correlated with the true exposure (eg, the larger the true exposure, the

[1] National Institute for Occupational Safety and Health, Cincinnati, Ohio, United States.
[2] Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio, United States.

Reprint requests to: Dr K Steenland, National Institute for Occupational Safety and Health (NIOSH), R-13, 4676 Columbia Parkway, Cincinnati, Ohio 45226, United States. [E-mail: kns1@cdc.gov]

larger the error). This situation is in contrast to the classical error model, in which the error is independent of the true exposures (and usually assumed to be normally distributed) (1),

$$\text{exposure}_{\text{observed}} = \text{exposure}_{\text{true}} + \varepsilon.$$

An example of classical measurement error would be when measurements are made on all workers in a study and exposure is assigned to each worker according to the measurements (eg, radiation levels taken from a radiation badge). Such complete data on exposure is uncommon in occupational epidemiology, particularly in retrospective studies.

It is well known that, in linear models with continuous disease outcomes, the effect of classical errors in measuring exposure causes a bias towards the null in the estimated effect of exposure (ie, in estimating exposure-response trends). It is also well known that in the case of Berkson errors the estimate of the exposure-response trend will be unbiased, although the variance of the estimate will be affected (1). Seixas et al (3) have discussed the assignment of job-specific arithmetic means (leading to Berkson errors) in the linear case, without taking into consideration duration of exposure. Armstrong (4, 5) provides useful reviews of measurement errors in general.

In log-linear models with dichotomous disease outcomes, unlike the linear case, it has been shown theoretically that Berkson measurement error can result in bias either towards or away from the null in estimating exposure-response trends (6, 7). Deddens & Hornung (7) showed that bias was generally away from the null when (i) the disease was rare, (ii) the true exposure level was distributed log-normally, and (iii) the error variance increased with increasing exposure level. These are common conditions in occupational epidemiology. Chronic diseases are frequently rare, and exposures are typically distributed log-normally. The third condition will also occur, for example, in a situation in which the variance of the job-specific exposure level is assumed to increase with the job-specific mean exposure level and the job-specific mean exposure levels from a sample of workers are used to assign exposure level to all workers within a specific job. A positive correlation between the job-specific variance of exposure level and the job-specific mean is common in occupational studies. For example, in silica exposure data across 133 different jobs in 4 industrial sand plants, where multiple measurements were taken per job, the correlation coefficient between the job-specific mean and job-specific variance was 0.68 (P=0.0001) (8).

Deddens & Hornung (7) considered only exposure level, without consideration of duration of exposure. The use of cumulative exposure (the sum of duration times the exposure level across all jobs) brings a second variable (duration) into the problem. The distribution of the assigned cumulative exposure then depends on the distribution of the duration of exposure, as well as the distribution of the exposure level. The duration of exposure in many occupational studies is known with little or no error; for example, it can be obtained from company personnel records. The observed duration of exposure can be reasonably assumed to be log-normally distributed, given that there is usually a long right-sided tail to the distribution with a few workers exposed for a long period of time.

We have conducted simulations to explore potential bias further in the estimate of exposure effect when cumulative exposure is the exposure metric of interest and when job-specific exposure levels are assigned to all workers based on job-specific means from a sample of workers.

Methods

Basic model

In each simulation we considered 30 000 study subjects, 10 000 in each of 3 jobs. The workers were assumed to have remained in the same job over time. Duration of exposure was assumed to be measured without error and to be log-normally distributed, with a mean of 0.5 years and a standard deviation ranging from 0.1 to 0.9, depending on the simulation.

For each simulation, we generated 30 000 durations and assigned them randomly to each worker.

True exposure level for each job was assumed to be distributed log-normally with means of 1, 2, or 3, with standard deviations increasing across jobs (eg, 0.7, 1.4, and 2.1, respectively). We assigned a true exposure level to each subject for each simulation by generating 10 000 exposure levels for each job and randomly assigning to the 10 000 workers in that job.

The true cumulative exposure for each subject was then created by multiplying the duration and true exposure level for each subject.

Disease status was then assigned to each subject. The assignment was done by assuming a true cumulative exposure-response trend for a dichotomous outcome, as well as a binomial (logistic) model. The exposure coefficient (β_1) was chosen arbitrarily to be 0.6. The intercept (β_0) was chosen (-4.0) to result in a rare disease (5%). The true cumulative exposure-response model for all the simulations was therefore:

$$\text{Log}\,(p/(1-p)|x) = -4.0 + 0.6^{*}x$$

or

$$p = 1/\{1 + \exp(4.0 - 0.6x)\},$$

where p was the probability of disease and x was the true cumulative exposure. Given the known true exposure, a probability of disease was then calculated for each subject. On the basis of the specific probability, each subject was randomly assigned either 1 or 0 (disease or nondisease) using an SAS function (RANBIN) (9).

We then constructed a mismeasured or "assigned" exposure level for each subject by choosing a random sample of 100 subjects from each job category and calculating the arithmetic mean of their true exposure level. This mean exposure level was then assigned to all the subjects in that job category. Finally, we calculated a mismeasured or "assigned" cumulative exposure for each subject by multiplying their mismeasured exposure level by their duration of exposure.

Logistic regression analyses were then conducted using either the true cumulative exposure or the assigned cumulative exposure. For either the true or the assigned cumulative exposure, each simulation was repeated 100 times, and the average regression coefficient and its observed standard error were reported for both true and assigned cumulative exposure.

Variations on the basic model

We varied some of the assumptions of the basic model and conducted further simulations as follows:

1.The exposure level variance was kept constant across jobs; some results are shown for this scenario.

2. The Cox regression was used instead of logistic regression. We generated exponential survival times, for which the mean or hazard depended on the true cumulative exposure with a coefficient of 0.6. We then censored the times using a uniform distribution to determine the overall frequency of disease occurrence. Cox regressions were carried out using either the true cumulative exposure or the assigned cumulative exposure. The results were similar to the logistic model and have not been presented here.

3. Exposure level was allowed to change over time, decreasing 10-fold over a 10-year period via a stepwise function and then remaining constant. True cumulative exposure then required summing over a different exposure level for each year a worker was on the job. The mean exposure level for a sample of 100 workers in each job was calculated by taking the mean of the cumulative exposure for each of the 100 workers divided by their duration in the job. The simulations yielded results similar to those obtained when exposure was assumed to be constant over time, and they have not been presented.

4. Disease was changed to be more common (30% prevalence); again the same pattern of results was observed, although the pattern was less pronounced; the results are not presented.

5. The job-specific geometric mean exposure level within each job was assigned, instead of the job-specific arithmetic mean. The results generally yielded regression coefficients for assigned cumulative exposure which were higher than those based on the arithmetic mean but which exhibited the same pattern of biases.

6. Job-specific exposure levels were normally rather than log-normally distributed. The same overall pattern of bias resulted, although it was somewhat less pronounced, and the results have not been presented.

Results

Table 1 gives the job-specific data for cumulative exposure (true and assigned) for a single simulation. The mean of the assigned cumulative exposure approximates the mean of the true cumulative exposure (depending on how well the sample of the exposure level for 100 workers per job, used to assign exposure level, represented all the workers in that job). The standard deviation of the assigned cumulative exposure was less than the true cumulative exposure, as expected.

Figure 1 shows the average regression coefficients for the cumulative exposure that resulted from the job-specific mean being assigned to all the subjects when the variation of exposure remained constant across jobs, across different standard deviations of duration. Three different standard deviations of the true exposure level were used (0.1, 0.5, 0.9) for these simulations. Regardless of the standard deviation of the duration or the standard deviation of the exposure level used, there was little bias (departure from the true regression coefficient of 0.6) in the estimated exposure coefficients for the cumulative exposure. This result would have been predicted from the earlier work by Deddens & Hornung (7). Assigning the geometric mean rather than the arithmetic mean also resulted in an unbiased regression coefficient for this scenario (results not shown).

Figure 2 shows the average regression coefficient which resulted from the mean exposure level being assigned to all the workers in a specific job when the standard deviation of the job-specific exposure-level increased proportionally with the job-specific mean. Data are shown for a range of different standard deviations of

Table 1. Mean job-specific true and assigned cumulative exposure for one simulation (30 000 workers, 10 000 per job, standard deviation of duration of exposure 0.7).

	True cumulative exposure		Assigned cumulative exposure	
	Mean	SD	Mean	SD
Job 1	0.4958	0.9186	0.4858	0.7259
Job 2	1.0062	1.8607	0.9369	1.2378
Job 3	1.4895	2.7583	1.6417	2.2810

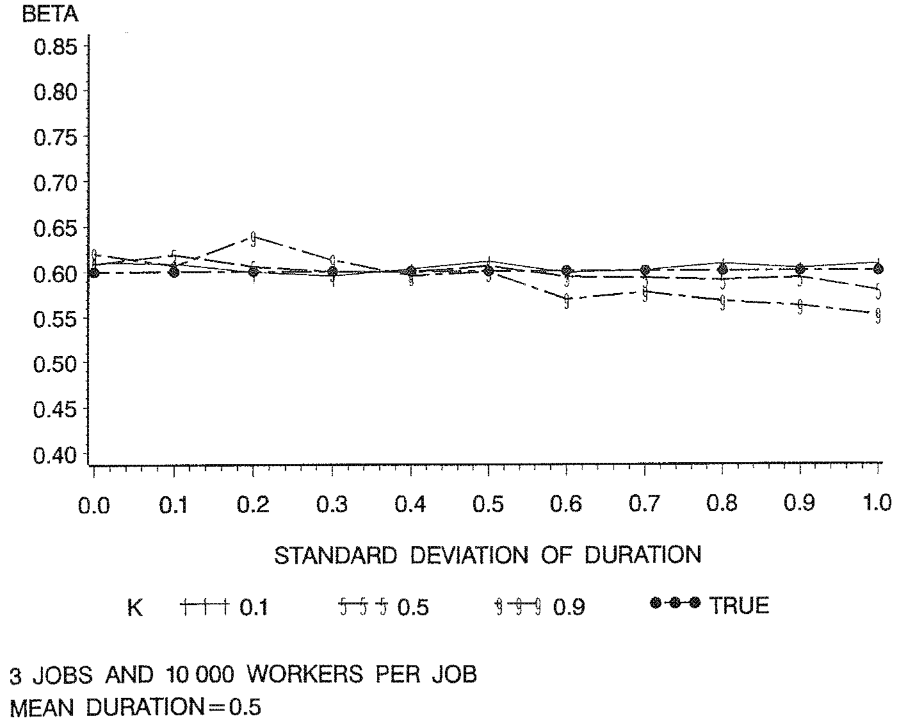

Figure 1. Average parameter estimates versus standard deviation (STD) of duration after the assignment of job-specific means to workers using 100 simulations, where K satisfies STD(exposure) = K.

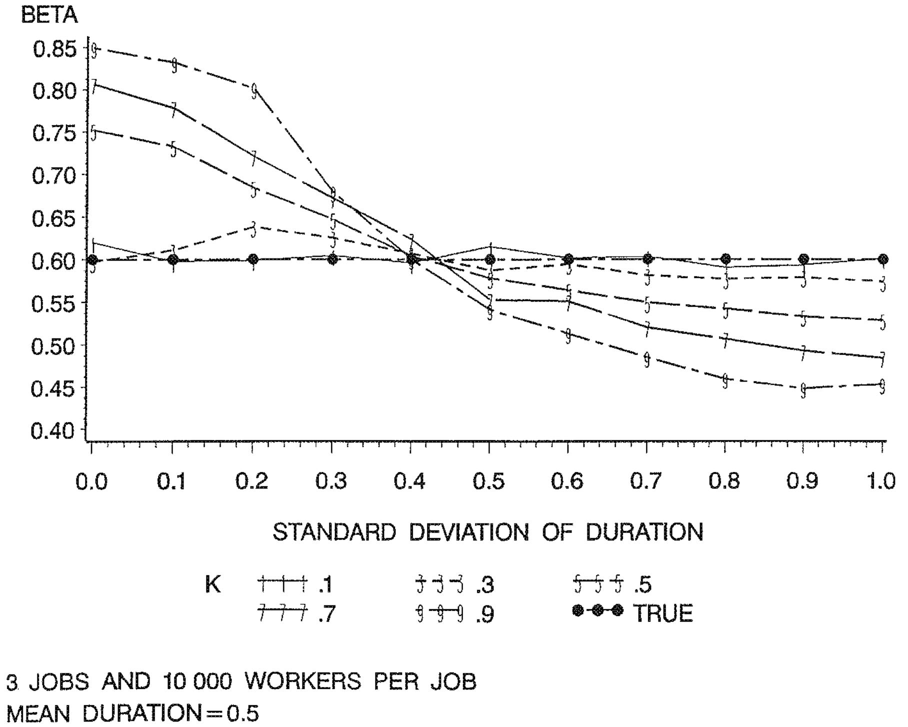

Figure 2. Average parameter estimates versus the standard deviation (STD) of duration after the assignment of job-specific means to workers using 100 simulations, where K satisfies STD(exposure) = K · mean(exposure).

duration. Five different exposure-level standard deviations were used for the lowest exposed job (0.1, 0.3, 0.5, 0.7, and 0.9), and in each case they increased proportionately for the other 2 jobs (eg, when 0.3 was used for job 1, 0.6 and 0.9 were used for jobs 2 and 3). This figure shows that, when exposure level has a high standard deviation (eg, ≥0.5), the regression coefficient is upwardly biased when the standard deviation of duration is low, and it is downwardly biased when it is high. The bias can be substantial (eg, 30—50%).

Table 2 gives partial results for the scenario used in figure 2, for an exposure-level standard deviation of 0.7

for the lowest exposed job and for 2 different standard deviations of duration. Table 2 shows that use of the true exposure in fact resulted in the expected regression coefficient for cumulative exposure (0.6). The same pattern seen in figure 2 is apparent; when the standard deviation of duration is small, assignment of the arithmetic mean gives overestimates, while the opposite is true when duration has a large standard deviation. Table 2 also shows that the assignment of the geometric mean of the exposure level resulted in higher estimates than did the assignment of the arithmetic mean. These estimates also tended to decrease as the standard deviation of exposure increased.

The estimated cumulative exposure-response trends in figure 2 are assumed (by the logistic model used) to be linear in the log odds. Figures 3 and 4 show the shape of the exposure-response curve (log odds versus cumulative exposure) via restricted cubic spline curves [5 knots at 5th, 25th, 50th, 75th, and 95th percentiles of exposure (11)]. The spline for figure 3, when duration had a low variance of duration, shows a consistently overestimated exposure-response trend, conforming to a linear model. However, the spline for figure 4, with a high duration variance, shows that the assignment of the job-specific means resulted in overestimates of risk in the middle range of cumulative exposure and underestimates of risk at the highest cumulative exposures; the overall linear estimate was biased downwards. The fact that measurement error led to an underestimation of risk in figure 4 in the highest range of cumulative exposure may be part

Table 2. Log-linear model. Disease as a function of cumulative exposure, log-normal exposure within job, 3 jobs with 10 000 people, 100 replications each simulation.[a]

Duration	True exposure		Exposure coefficients					
	Mean	SD	True exposure		Assigned exposure		Assigned exposure	
			Mean	SD	Arithmetic mean	SD	Geometric mean	SD
Mean 0.5, SD 0.1	1	0.7 [c]	0.602	0.002	0.778	0.009	0.945	0.011
	2	1.4 [d]	.	-	-	-	-	-
	3	2.1 [e]	.	.	.	.	.	.
Mean 0.5, SD 0.7	1	0.7 [c]	0.599	0.001	0.523	0.003	0.636	0.004
	2	1.4 [d]	.	.	.	.	.	.
	3	2.1 [e]	.	.	.	.	.	.

[a] Simulations conducted via logistic regression, with β_0=-4.0, β_1=0.6. [b] Observed standard error of the mean regression coefficient across the 100 simulations. [c] Job 1. [d] Job 2. [e] Job 1.

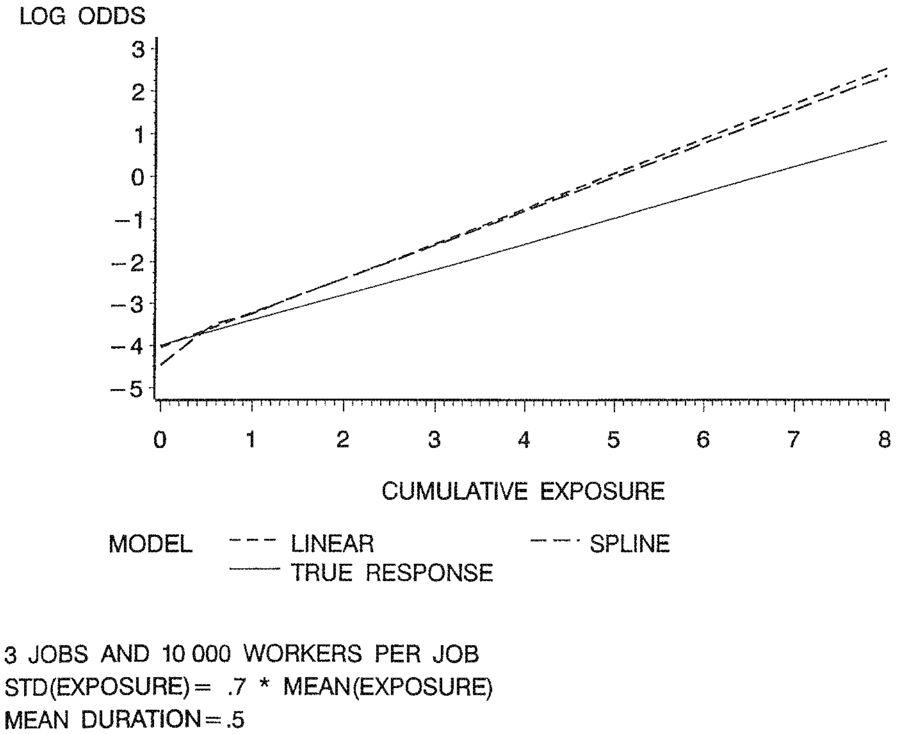

Figure 3. Linear and spline estimates of exposure response after the assignment of job-specific means to workers in different jobs (standard deviation = 0.1 for duration).

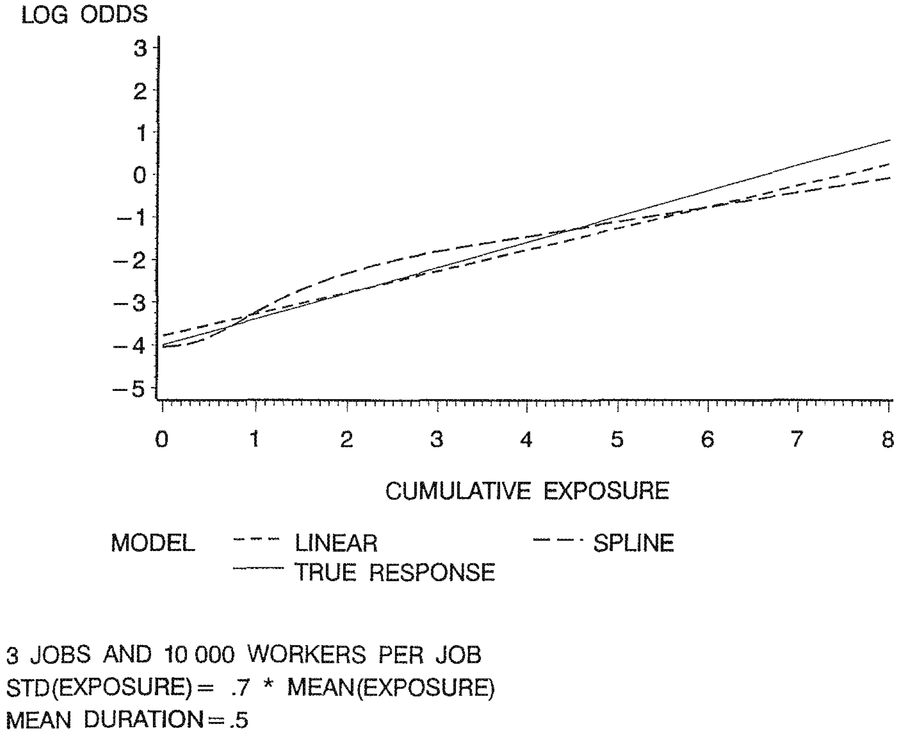

Figure 4. Linear and spline estimates of exposure response after the assignment of job-specific means to workers in different jobs (standard deviation = 0.9 for duration).

of the reason why relative risks in occupational studies sometimes tend to tail off at the highest exposures (eg, see references 11, 12). The distortion of the true shape of the exposure-response curve, as seen in figure 4 (eg, from linear to curved), resulting from "multiplicative" error (ie, error proportional to the mean) when the exposure level is lognormal, has been noted previously by other authors (4).

Discussion

The assignment of job-specific mean exposure levels to all workers in a job, resulting in Berkson-type errors, is common in occupational epidemiology. With log-linear models, as seen in figure 1, the assignment of the arithmetic mean resulted in an unbiased estimate of the exposure-response trend (for cumulative exposure) when the variation of the exposure level was constant across jobs. However, we believe this situation is uncommon in occupational settings. For the more common scenario in which the variance of exposure level across jobs increases with job-specific means, regression coefficients estimating cumulative exposure-response trends were biased upwards when the variance of duration was low. This result would be expected based on the work of Deddens & Hornung (7), in which upward bias was shown when duration was not taken into account in the exposure metric (ie, only exposure level was considered); this is tantamount to using cumulative exposure but assuming that duration is the same for all workers (ie, the variance of duration is 0). As the variation in duration increased in our data, however, the estimated regression coefficients based on assigned exposure level were biased downwards. When the standard deviation of duration was approximately equal to its mean, the regression coefficients for cumulative exposure were approximately unbiased.

To investigate the distribution of duration further, we considered 3 reasonably typical industrial cohorts (dioxin, industrial sand, ethylene oxide). We found mean durations of employment of 12, 8, and 9 years, while the respective standard deviations were 12, 10, and 9 years. This result would suggest that, in actual data, the standard deviation of duration may not differ greatly from its mean. In these 3 cohorts the distribution of duration was skewed, so that the data indeed approximated a log-normal distribution, as was assumed in our simulation.

It should be noted that the effects of measurement error, as observed in this work, were negligible when the standard deviation of the job-specific exposure level was low (ie, less than half the job-specific mean). However, some empirical data suggest that this is rarely the case. Job-specific exposure data on silica in industrial sand plants, for example, suggests, on the average, that the job-specific standard deviation of exposure level is approximately equal to its mean (8). Occupational and environmental exposure data often show wide variance.

It should also be noted that our results are largely independent of scale. For example, we could have used a larger mean duration in our simulations, which would have resulted in a larger cumulative dose, which in turn would result in a smaller regression coefficient for cumulative dose if it is assumed that the percentage of diseased subjects in our study population was kept constant.

As noted earlier, the same pattern of results holds when we allowed exposure level to change over time, when we used a Cox regression instead of a logistic regression, when we allowed a higher disease prevalence (30% instead of 5%), when we assumed exposure to be normally distributed instead of log-normally distributed, and when we assigned the geometric mean instead of the arithmetic mean (although we do not recommend the geometric mean, as the estimated regression coefficients are consistently higher than with the arithmetic mean, without any reduction in bias). While the same patterns resulted, they were less pronounced when exposure was normally distributed or when the disease was less rare.

Despite the preceding sensitivity analyses in which we varied some of the parameters of our simulations, we note that our work is directly applicable only to the conditions we simulated, and these conditions were necessarily limited. Simulations are only as good as their assumptions (13). Our assumptions refer to a restricted region of the total "parameter space" which could be considered. Nonetheless we have tried to use reasonable assumptions which conform to actual data seen in occupational studies.

In summary, estimated trends of disease risk by cumulative exposure can be biased in studies in which data from a sample of workers are used to assign exposure levels to all workers. The direction of the bias depends on the variation of the duration, with increasing variation leading to downward bias. Fortunately, it appears that, under reasonably common conditions in which the duration of exposure is approximately equal to its standard deviation, regression coefficients are approximately unbiased. The underlying reasons for these patterns remain unclear; the distribution of cumulative exposure, as simulated in this report, is complex, and theoretical results to support the observed pattern are likely to be hard to derive.

References

1. Snedecor G, Cochran W. Statistical methods. Ames (IA): Iowa State University Press, 1989.
2. Berkson J. Are there two regressions? J Am Stat Assoc 1950; 45:164—80.
3. Seixas N, Robins T, Mouton L. The use of geometric and arithmetic mean exposure in occupational epidemiology. Am J Ind Med 1988;14:465—77.
4. Armstrong BG. Effect of measurement error on epidemiological studies of environmental and occupational exposures. Occup Environ Med 1988;55:651—6.
5. Armstrong BG. Effects of measurement errors on relative risk regressions. Am J Epidemiol 1990;132:1176—84.
6. Prentice R. Covariate measurement errors and parameter estimation in a failure time regression model. Biometrika 1982; 69:331—41.
7. Deddens J, Hornung R. Quantitative examples of continuous exposure measurement errors that bias risk estimates away from the null. In: Smith C, Christiani D, Kelsey K, editors. Chemical risk assessment of occupational health. London: Auburn, 1994:77—85.
8. Amandus H. A feasibility study of the adequacy of company records for a proposed NIOSH study of silicosis in industrial sand workers. Morgantown (WV): National Institute for Occupational Safety and Health (NIOSH), Jan 30, 1990. NIOSH report, NIOSH DRDS.
9. SAS. SAS user's guide: statistics (version 6.07). Cary (NC): SAS Institute, 1991.
10. Harrell F, Lee K, Pollack B. Regression models in clinical studies: determining relationships between predictors and response. JNCI 1988;80:1198—202.
11. Steenland K, Deddens J, Stayner L. Diesel exhaust and lung cancer in trucking industry: exposure-response analyses and risk assessment. Am J Ind Med 1998;34:220—8.
12. Stayner L, Smith R, Bailer A, Gilbert S, Steenland K, Dement J, et al. An exposure-response analysis of respiratory disease risk associated with occupational exposure to chrysotile asbestos. Occup Environ Med 1997;54:646—52.
13. Maldonado G, Greenland S. The importance of critically interpreting simulation studies. Epidemiology 1997;8:453—6.

Received for publication: 21 December 1998